Prediction of genomic breeding values for dairy traits in Italian Brown and Simmental bulls using a principal component approach

PINTUS, M. A; GASPA, GIUSTINO; NICOLAZZI, EL; VICARIO, D; ROSSONI, A; AJMONE MARSAN, P; NARDONE, A; DIMAURO, CORRADO; MACCIOTTA, NICOLO' PIETRO PAOLO,

Prediction of genomic breeding values for dairy traits in Italian Brown and Simmental bulls using a principal component approach,

JOURNAL OF DAIRY SCIENCE, 95: 3390: 3400, 2012,
doi: 10.3168/jds.2011-4274

1

2 **Interpretive Summary**

3

4 **Prediction of Direct Genomic Values for dairy traits in Italian Brown and Simmental Bulls**
5 **using a Principal Component Approach** *By Pintus et al.*

6 In this work, principal component analysis is used to reduce the number of predictors for calculating
7 direct genomic breeding values (DGV) for bulls of two cattle breeds in Italy. The PC method allows
8 for a relevant reduction (about 94%) in the number of independent variables when predicting DGV,
9 with a huge decrease in calculation time and without losses in accuracy compared to the direct use
10 of SNP genotypes.

11
12

Prediction of Direct Genomic Values for dairy traits in Italian Brown and Simmental Bulls using a Principal Component Approach.

M.A Pintus[*], G. Gaspa[*], E.L. Nicolazzi[§], D. Vicario[**], A. Rossoni[§§], P. Ajmone-Marsan[§], A. Nardone[†], C. Dimauro[1], N.P.P. Macciotta[*,1]

*Dipartimento di Scienze Zootecniche, Università di Sassari,Sassari, Italy, 07100.

§Istituto di Zootecnica, Università Cattolica del Sacro Cuore, Piacenza, Italy, 29100.

**Associazione Nazionale Allevatori Razza Pezzata Rossa Italiana (ANAPRI), Udine, Italy, 33100.

§§Associazione Nazionale degli Allevatori di Razza Bruna (ANARB), Verona, Italy, 37012.

†Dipartimento di Produzioni Animali, Università della Tuscia, Viterbo, 01100.

[1]Corresponding author: Nicolò P.P. Macciotta, Dipartimento di Scienze Zootecniche, Università di Sassari, via De Nicola 9, 07100 Sassari, Italy. Phone number: 0039 079229298. Fax number: 0039 079229302. e-mail: macciott@uniss.it

ABSTRACT

The huge number of markers in comparison with phenotypes available represents one of the main issues in genomic selection. In this work, principal component analysis (PCA) was used to reduce the number of predictors for calculating direct genomic breeding values (DGV) and genomic enhanced estimated breeding values (GEBV). Bulls of two cattle breeds in Italy (749 Brown and 479 Simmental) were genotyped with the 54K Illumina beadchip. After data editing, 37,254 and 40,179 SNP were retained for Brown and Simmental, respectively. Principal component analysis carried out on SNP genotype matrix extracted 2,257 and 3,596 new variables in the two breeds, respectively. Bulls were sorted by birth year or randomly shuffled to create reference and prediction populations. The effect of principal components on de-regressed proofs in reference animals was estimated with a BLUP model. Results were compared to those obtained by using SNP genotypes as predictors either with BLUP or Bayes_A methods. Traits considered were milk, fat and protein yield, fat and protein percentage, somatic cell score, and udder score. GEBV were obtained for prediction population by blending DGV and PA. No substantial differences in correlations between DGV and EBV were observed among the three methods in the two breeds. The approach based on the use of PCA showed the lowest prediction bias. The PCA method allowed for a reduction of about 90% in the number of independent variables when predicting DGV, with a huge decrease in calculation time and without losses in accuracy.

Key words: SNPs, genomic selection, principal component analysis, accuracy

INTRODUCTION

Advancements in genome sequencing technology have been implemented into high throughput platforms able to genotype simultaneously tens of thousands SNP markers distributed across the whole genome of livestock species (Van Tassell et al., 2008). Dense marker maps are nowadays used in cattle breeding for genome-wide association studies (Cole et al., 2009, Price et al., 2006) and for predicting genomic-enhanced breeding values (GEBV) of candidates to become sires and dams in genomic selection (GS) programs (Meuwissen et al., 2001). The basic framework of genomic selection involves two steps. Firstly, effects of chromosomal segments are estimated in a set of reference animals with known phenotypes and SNP genotypes. Then estimates are used to predict Direct Genomic Values (DGV) of animals for which only marker genotypes are known. DGV are usually blended with other measures of genetic merit such as official pedigree index (PI) to obtain the final GEBV (Ducrocq and Liu 2009; VanRaden et al., 2009). GS programs have already been implemented in different countries to evaluate young bulls entering progeny testing, achieving reliabilities higher than those of PI (Hayes et al., 2009a, VanRaden et al., 2009). Expected benefits of the GS are the reduction of generation intervals, the increase of EBV accuracies for females and a cost reduction for progeny testing (Konig et al., 2009, Schaeffer, 2006).

However, several issues are still to be addressed in GS. Examples are the assessment of the frequency of marker effect re-estimation along generations (Solberg et al., 2009), the evaluation of the impact of population structure on estimated effects (Habier et al., 2010), and the choice of the most suitable mathematical model and dependent variable for the estimation step (Guo et al., 2010). Apart from situations in which the number of genotyped animals is quickly approaching or overcoming the number of markers used, as the North American genomic project (VanRaden and Sullivan, 2010), the huge imbalance between predictors and observations still represents the main constraint to the implementation of GS programs, especially for breeds other than Holstein.

A way to reduce this data asymmetry could be found in combining data from different populations of the same breed or from different breeds in a common reference set, both within and across countries (Boichard et al, 2010). Reports on simulated and real data show some increases in DGV accuracies, but results are strongly dependent on the genetic similarity between breeds and/or on the trait analyzed (de Roos et al., 2009, Hayes et al., 2009b).

A different strategy is based on the reduction of the number of predictors used in the estimation equations. A straightforward approach is to perform a preliminary selection of markers on the basis of their relationship with the phenotype or of their chromosomal location (Hayes et al., 2009a, Moser et al., 2010, Vazquez et al., 2010). An alternative is represented by the Bayes_B method that retains markers with non-zero effect on phenotypes directly during the estimation step (Meuwissen et al., 2001, VanRaden, 2008). Other approaches of SNP selection have been proposed mainly for genome-wide association analyses (Aulchenko et al., 2007, Gianola et al., 2006, Gianola and van Kaam, 2008, Long et al., 2007). In all the above mentioned methodologies, SNP selection is based on their relevance to the considered phenotype. Thus specific sets of markers may be required for different traits.

An alternative to marker selection for reducing predictor dimensionality is represented by their synthesis via multivariate reduction techniques. In particular, principal component analysis (PCA) and Partial Least Squares Regression (PLSR) have been proposed for estimating DGV (Solberg et al., 2009). Actually, in the PLSR approach the extraction of latent variables from predictors is carried out by maximizing their correlation with the dependent variable(s). Thus the reduction of the system dimension is still based on the magnitude of the predictor effects on the considered trait. On the contrary, PCA is entirely based on the factorization of the SNP (co)variance (or correlation) matrix. This technique allows for a huge reduction of the number of independent variables (>90%) in the estimation of DGV while achieving accuracies comparable to those obtained using all SNP genotypes (Macciotta et al., 2010, Solberg et al., 2009). A recent

comparison highlighted good accuracies of both dimension reduction techniques in predicting DGV for milk yield in US Holsteins (Long et al., 2011). Compared to other approaches of predictor reduction, PCA limits the loss of information because each SNP is involved in the composition of each principal component Moreover, extracted principal components are orthogonal, thus avoiding multicollinearity problems. The PCA approach also allows to model the variance structure of predictors in the BLUP normal equations by using eigenvalues as variance priors (Macciotta et al., 2010). Furthermore, PCA has been used in genome-wide association studies to reduce the number of dependent variables (Bolormaa et al., 2010).

The reduction of predictor dimensionality is a straightforward strategy when implementing GS with reference populations of limited size. This situation may occur in minor cattle breeds or in larger populations at early stages of GS programs. This is the case of the SELMOL project recently started in Italy that involves different cattle breeds (both dairy and beef).

Aim of this study is to calculate genomic breeding values for dairy traits in populations of limited sizes of Italian Brown and Simmental bulls by using the principal component approach for reducing the number of predictors. The PCA based method is compared with other approaches that fit directly all SNP genotypes available as predictors.

MATERIALS AND METHODS

Data

A total of 775 Italian Brown and 493 Italian Simmental bulls were genotyped at 54,001 SNP loci with the Illumina Bovine SNP50TM bead-chip. Considering the limited size of the sample, the priority in the edit was to keep the number of bulls as large as possible. A stringent selection was performed on markers. Edits were based on the percentage of missing data (<0.025), Mendelian inheritance conflicts, absence of heterozygous loci, minor allele frequency (<.05), deviance from Hardy-Weimberg equilibrium (<0.01) (Wiggans et al., 2009). Edits on animals were based on the

number of missing genotypes (<1,000) and on inconsistencies in the Mendelian inheritance (96 and 70 father-son pairs were included in the archives for Italian Brown and Simmental, respectively). An overall accuracy higher than 99% was obtained by double-genotyping some animals. A summary of the initial and final number of bulls and SNP, together with the impact of the different elimination steps is reported in table 1. In the final data, missing genotypes (in general less than the 0.5%) were replaced by the means of the observed genotypes at that specific locus.

Phenotypes used were both MACE de-regressed proofs (DRPF) provided by the two breed associations. Traits considered were milk, fat and protein yield (kg), fat and protein percentages, somatic cell score. Average reliabilities of DRPF were 0.87 (±0.08) and 0.92 (±0.04) for Italian Brown and Simmental bulls, respectively.

Animals were sorted by year of birth and the dataset split into reference (REF) and prediction (PRED) subsets, comprising older and younger animals, respectively. Three ratios of REF-PRED animals were considered (0.70:0.30, 0.80:0.20, 0.90:0.10). The distribution of years of birth in the different breeds is depicted in figure 1.

A common strategy when dealing with a small population of genotyped animals is to obtain different sets of reference and prediction by randomly picking up animals from the original archive (Luan et al., 2009). Thus, in the present study, PRED population (30% of animals) was also generated by extracting bulls at random from the 50% of youngest animals. Ten replicates were performed for each trait.

Statistical Models

Principal component analysis was used to extract latent variables from the SNP data matrix **M** with *m* rows (m= number of individuals in the entire data set, i.e. REF plus PRED) and *n* columns (n=number of SNP retained after edits). Each element (i,j) corresponded to the genotype at the j^{th} marker for the i^{th} individual. Genotypes were coded as -1 and 1, for the two homozygotes,

and 0 for the heterozygote, respectively. PCA was performed separately for each chromosome. On simulated data, analyses carried out either on the whole genome simultaneously or separately by chromosome did not affect DGV accuracy (Macciotta et al., 2010). PCA was carried out on the whole data set (REF+PRED) separately for each breed. The number of principal components retained (k) was based on the sum of their eigenvalues. An empirical threshold of 80% of explained variance was fixed according to indications of other authors (Boolorma et al., 2010). Scores of the selected components were calculated for all individuals.

For each breed, the estimation of predictor effects on the REF data set was carried out using the following BLUP model (PCA_BLUP):

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Zg} + \mathbf{e}$$

where **y** is the vector of DRPF, **1** is a vector of ones, μ is the general mean, **Z** is the matrix of PC scores, **g** is the vector of PC regression coefficients treated as random, and **e** is the vector of random residuals. Covariance matrices of random PC effects (**G**) and residuals (**R**) were modeled as diagonal $\mathbf{I}\,\sigma_{aj}^2\lambda$ and $\mathbf{I}\,\sigma_e^2$ respectively. In particular, the contribution of each j-th principal component to the genetic variance was assumed to be proportional to its corresponding eigenvalue (λ), i.e. $\sigma_{ji}^2 = (\sigma_a^2/k) * \lambda_j$ (Macciotta et al., 2010). Variance components were those currently supplied by breed associations for Interbull evaluations (http://www-interbull.slu.se/national_ges_info2/framesida-ges.htm). BLUP solutions were estimated using Henderson's normal equations (Henderson, 1985) solved by using a LU factorization where the left hand side part of mixed model equations was decomposed into the product of a lower and a upper triangular matrix, respectively (Burden and Faires, 2005).

To evaluate the effect of the PCA reduction of predictors on DGV accuracy, the estimation step was carried out also with two methods that fit all available SNP genotypes, but with different assumptions on the distribution of their effects.

176 The first was the BLUP (SNP_BLUP) method that assumed an equal contribution of each
177 marker locus to the variance of the trait, sampled from the same normal distribution (Meuwissen et
178 al., 2001). In this case, **Z** was the matrix of SNP genotypes coded as 0, 1 and 2. Mixed model
179 equations were solved using a Gauss-Seidel iterative algorithm.

180 The second was the Bayes_A method, that allowed for variance to differ across chromosome
181 segments on the assumption that a large number of SNP have small effects and few have a large
182 effect (Meuwissen et al., 2001). The fitted model (BAYES_A) was:

183 $$\mathbf{y} = \mathbf{1}\mu + \mathbf{Zg} + \mathbf{Wu} + \mathbf{e}$$

184 where **u** is a vector of polygenic breeding values assumed to be normally distributed, with
185 $u_i \sim N(0, \mathbf{A}\sigma_a^2)$, where **A** is the average relationship matrix and σ_a^2 is the additive genetic variance.
186 Prior structure and hyper-parameters were chosen according to Meuwissen *et al.* (2001). A scaled
187 inverted chi-squared prior distribution was assumed for SNP specific variances, under the
188 hypothesis that most of markers have nearly zero effects (i.e. markers not linked to any QTL) and
189 only few have large effects. A total of 20,000 iterations were performed, discarding the first 10,000
190 as burn-in and considering no thinning interval. A residual updating algorithm was implemented to
191 reduce computational time (Legarra and Misztal, 2008).

192 The general mean (μ) and the vector ($\hat{\mathbf{g}}$) of the principal component or marker effects
193 estimated either with BLUP (SNP_BLUP) or Bayes A (BAYES_A) in the REF population were
194 used to calculate the DGV for the jth animal in the PRED subset for each breed as:

195 $$\mathrm{DGV_j} = \mu + \sum_{i=1}^{k} \mathbf{z}'_{ij}\, \hat{\mathbf{g}}_i$$

196 where **z** is the vector of component scores or marker genotypes and *k* is the number of principal
197 component or markers used in the analysis. Pearson correlations between DGV and DRPF in PRED
198 individuals were calculated.

199 DGV obtained with three different estimation methods were blended with PI to obtain
200 GEBV by using the EDC as weighting factors:

201 $$\mathrm{GEBV_i} = \mathrm{DGV_i} \cdot edcG + \mathrm{PI_i} \cdot edc_i$$

202 where *edcG* and *edc* are the equivalent daughter contributions for DGV or PI respectively.
203 Values of *edcG* were calculated from the approximate DGV reliabilities, obtained as REL_{DGV} =
204 $(r^2_{DGV,DRPF})/REL_{DRPF}$ (Hayes et al., 2009), as

205 $edcG = k * REL_{DGV}/(1-REL_{DGV})$

206 where $k=(4-h^2)/h^2$. The same approach was used to calculate *edc* for PI. The procedure
207 followed was the same used to validate the international GEBV of Italian Simmental approved in
208 November 2011 (http://www.interbull.org).

209 Finally, in order to evaluate the efficiency of genomic predictions versus the traditional
210 polygenic evaluations, squared correlation between Genomic Enhanced estimated Breeding Values
211 and EBV ($R^2_{EBV\text{-}GEBV}$) were computed and compared with those between PI and EBV. Bias was
212 assessed by evaluating the regression coefficient of EBV on predicted GEBV.

213

214 **RESULTS**

215 A common criterion for choosing the number of principal components to retain is the visual
216 inspection of their eigenvalue pattern. As an example, Figure 2 reports the chromosome-wide
217 variance explained by each successive component extracted from SNP located on BTA6 in the
218 Brown breed. The eigenvalue was small also for the top two components (about 7% and 5%,
219 respectively) with a smooth decrease followed by a plateau reached at about 100 PCs (86% of
220 variance explained) for this chromosome. The number of retained principal components genome-
221 wide was 3,596 and 2,257 for the Simmental and Brown breeds, respectively. A similar amount of
222 components was retained by Long et al. (2011). In any case, a large reduction of predictor
223 dimensionality (less than 10% of the number of original variables) was realized.

The extracted principal components were able to distinguish Brown from Simmental bulls. Individual scores of the first principal component of BTA6, for example, discriminated the two breeds whereas the third component highlighted a larger heterogeneity within the Brown sample (Figure 3). In PCA, the meaning of each extracted component is usually inferred by looking at eigenvector coefficients, i.e. the weights of each original variable (in this case the SNP genotype) in the component. However, it would be very hard to achieve an interpretation by examining thousands of correlations. The meaning of extracted variables could be assessed indirectly by looking at their relationships with other characteristics of the considered individuals. For example, the third principal component for BTA6 in the Brown breed was negatively correlated with the observed average individual heterozygosity (-0.43) and its score average showed a progressive decrease across year of birth of bulls. Such an ability of PCA to cluster individuals based on causes of variation of SNP genotype frequency was reported also for simulated data (Macciotta et al., 2010).

Correlations between DGV and DRPF for PRED bulls in different scenarios are reported in table 2 for the two breeds. In general, DGV accuracies were low to moderate, as expected due to the reduced size of the reference populations considered. Small differences across estimation methods were found. PC_BLUP and BAYES_A performed generally better than SNP_BLUP, and especially for Italian Brown. PC_BLUP accuracies were similar or slightly higher than those of BAYES_A for yield traits, especially milk (on average +5% and +0.5% for Brown and Simmental, respectively). The Bayesian method performed better in the case of SCS (average differences of 12.8% and 9.1% for Brown and Simmental, respectively).

Table 3 reports DGV accuracies for milk yield in the two breeds, obtained by creating PRED data by randomly picking up bulls from the 50% youngest animals. For brevity, only results for the PCA_BLUP approach are reported. Accuracies tended to increase, sometimes markedly, as in the case of somatic cell count for Brown, (+12.6% and +8.3% for Brown and Simmental,

249 respectively). These results do not agree with previous reports of Luan et al (2009) for Norwegian
250 Red Bulls, who did not find substantial differences in DGV accuracies of PRED animals obtained
251 by randomly shuffling the original data set or by sorting bulls according to their progeny testing
252 year. In the present work, similar improvement of DGV accuracies were observed for all statistical
253 approaches.

254 Squared correlations between GEBV or PI and EBV are reported in table 4 and 5 for the two
255 breeds. $R^2_{GEBV,DGV}$ values for Brown were substantially lower than those found for Simmental,
256 except for fat and protein percentages that showed opposite behavior. Squared correlations of
257 pedigree indexes were generally lower than those for GEBV in the Brown breed. Similar behavior
258 could also be observed for the Simmental, except for whereas higher except for fat and protein
259 percentages. PC_BLUP and BAYES_A gave better performances compared to the SNP_BLUP
260 method in Brown bulls. Finally, enlarging the ratio REF:PRED size seemed to increase $R^2_{EBV,GEBV}$
261 in Brown whereas no effect have been observed in Simmental.

262 In particular, squared correlations ranged from 0.01 to 0.39 for Italian Brown (Table 4).
263 Lowest values were obtained for yield traits, in particular for milk and protein (on average <0.1).
264 Highest $R^2_{EBV,GEBV}$ were observed for fat percentage, protein percentage, and somatic cell count
265 (on average 0.35, 0.32 and 0.15, respectively). Olson et al. (2011) reported the same value of
266 genomic prediction accuracy for SCS in a study on 1,188 brown Swiss bulls. These authors
267 observed higher values for yield traits. Accuracies for protein percentages reported in Table 4 agree
268 with results obtained on Australian Holsteins and Jerseys using different approaches and a
269 comparable size of reference population (Hayes et al., 2009; Moser et al., 2009). Best results in
270 genomic predictions for protein percentage have been also observed on US Holsteins (VanRaden et
271 al., 2009).

272 $R^2_{EBV,GEBV}$ obtained for the Simmental bulls ranged from 0.05 to 0.37 (Table 5). Values for
273 milk yield were on average (0.35 across all scenarios and methods) about five times compared to

274 the Brown breed. Yield traits had higher values compared to composition traits. For some scenarios,
275 squared correlations for protein yield were similar to those recently reported for Fleckvieh cattle
276 (Gredler et al., 2010). Intermediate values accuracies were obtained for somatic cell count (0.20on
277 average). PC_BLUP and BAYES_A slightly outperformed the SNP_BLUP approach. As in the
278 case of Brown, PC_BLUP gave slightly larger values than BAYES_A for yield traits and smaller
279 for composition traits, respectively. A part from fat and protein percentage, $R^2_{EBV,GEBV}$ were higher
280 than $R^2_{EBV,PI}$ for all estimation methods.

281 Regression coefficients of EBV on Genomic enhanced estimated breeding values (Table 6)
282 showed variability across breeds, methods, and traits. Differences between breeds were evident for
283 yield traits, with lower values for Brown bulls. For these traits, regression slopes were quite close to
284 the unity for all of the three methods and for all scenarios in the Simmental breed. For composition
285 traits and SCS, regression coefficients were lower than one indicating an underprediction of EBV
286 for high values and overprediction for low values. An opposite behavior can be opbserved for
287 Brown. The PC_BLUP method showed the lowest variability across traits.

288

289 **DISCUSSION**

290 In this paper, direct genomic breeding values genomic enhanced estimated breeding values
291 for some dairy traits were estimated by reducing the dimensionality of predictors with the principal
292 component analysis. Such a reduction aimed at simplifying data handling and at reducing
293 computational burdens while retaining most of the information. The PCA approach was compared
294 with some of the most popular methods used to predict DGV and GEBV, i.e. BLUP regression and
295 Bayes A, that fits directly all marker genotypes available but with different theoretical assumptions
296 on the distribution of their effects.

297 The BLUP methodology overcomes formally the problem of degrees of freedom in the
298 estimation step by fitting SNP effects as random rather than fixed (Meuwissen et al., 2001; Muir,

2007). However, the curse of dimensionality still represents the most important theoretical constraint for GS implementation. This problem is enhanced when a small number of genotyped animals is available, as in the case of this study. Actually, PCA does not completely address such an issue because of the data structure. The SNP correlation matrix is singular and therefore the number of eigenvalues different from zero is equal to the number of animals (i.e. the rows) minus one (Bumb, 1982; Patterson et al., 2006). In this study, PCA was carried out separately by each chromosome. At this level, the gap between predictors and observations was reduced and the number of components retained per chromosome (on average 75 and 120 in Brown and Simmental, respectively) was markedly smaller than the number of markers and of animals.

In agreement with previous findings on both simulated and real data, PCA was able to efficiently describe the correlation matrix of SNP genotypes (80% of explained variance) with less than 10% of the original variables. Such a reduction had a straightforward impact on calculation time. The PC_BLUP approach required about 2 minutes using a personal computer with a 2.33 GHz Quad core processor and 3.25 Gb of RAM. On the other hand, 6 to 9 hours were needed on average for the SNP_BLUP and Bayes_A approaches using a Linux server with 4 x 4 quad core processors and 128 Gb RAM. PCA required approximately half an hour, but it had to be done just once at the beginning of the work. Although calculation speed is not usually considered a technical priority for GS, compared for example to genotyping costs, it is likely to become more relevant due to the recent development of a larger (800K) SNP platform and to the upcoming very low cost sequencing technologies (Van Raden et al., 2011).

Of great interest is that such a huge reduction of calculation time did not result in a loss in DGV GEBV accuracy. The similarity of results between the PC_BLUP approach and the other two methods considered in the present paper confirms previous findings obtained with another multivariate dimension reduction technique, the Partial Least Squares Regression (Long et al., 2011; Moser et al., 2010, Moser et al., 2009). The reduction of the predictor dimensionality

obtained by selecting subsets of SNPs based on their chromosomal location or on their relevance to the trait usually resulted in a decrease of DGV accuracy (VanRaden et al., 2009, Vazquez et al., 2010). Actually, compared to subset SNP selection, the multivariate reduction has the advantages of not discarding any marker and of using uncorrelated predictors. The latter feature is confirmed by the observed lower bias of the PCA method compared to the SNP_BLUP method.

The similar results obtained by using methods characterized by different theoretical foundations suggests further considerations. The BLUP assumption of an equal effect of all markers on the variance of the trait is commonly considered rather inadequate to fit the assessed distribution of QTLs, i.e many loci with a small effect and a few with large effects (Hayes and Goddard, 2001). On the other hand, the superiority of the Bayesian approach that fits heterogeneous variances across chromosome segments is marked in simulations but not in real data; (Hayes et al., 2009a, Moser et al., 2009, VanRaden et al., 2009). Genome-wide association studies on human height suggest that genetic variation is explained by many loci of small additive effects (Yang et al., 2010). Moreover, a superior predicting ability of GEBVs for models that assume a heavy-tailed distribution of gene effects compared with finite locus models has been recently reported (Cole et al., 2009). Thus also BLUP methodology, even though not very accurate in terms of description of gene effect distribution, may offer robust DGV estimates (Goddard, 2009) with reasonable accuracies.

A possible criticism to the use of PCA is the lack of biological meaning of the extracted variables. Such a feature is in contrast with the general aims of the use of molecular markers in animal breeding, i.e. the overcome of the black-box approach of traditional quantitative genetics. However, even though a clear interpretation based on eigenvectors is not feasible, some results obtained in this work are worth to be mentioned. The extracted PC scores have been able to cluster animals of the two breeds, confirming the ability of this statistical technique to capture genetic variation across and within populations, highlighted in human genetic studies (Cavalli-Sforza and Feldman 2003, Paschou et al., 2007; Price et al., 2006). Moreover, a relationship between one of the

extracted PC and the average individual heterozygosity has been evidenced. It is interesting to notice that, in the case reported for BTA6, it was not the first extracted component to show the relationship with heterozygosity but the third one. This is also a distinguishing common feature of PCA: the first extracted component seldom contains biologically relevant information whereas these may be retrieved in components associated to smaller eigenvalues (Jombart et al., 2009).

In general, DGV accuracies GEBV reliability were rather low, as expected due to the reduced size of the sample of bulls considered and to their distribution across years of birth. Composition traits and udder score showed higher accuracies compared to yield traits,. These results, in agreement with previous findings (Hayes et al., 2009a, VanRaden et al., 2009), may reflect some variation in the genetic determinism of the trait (Cole et al., 2009). In particular, genes with large effects for fat and protein percentages have been discovered (Cohen-Zinder et al., 2005, Cole et al., 2009, Grisart et al., 2002). Thus, considering that genomic predictions works by tracking the inheritance of causal mutations (VanRaden et al., 2009), the method may be more efficient for traits where few loci affect a large proportion of the genetic variance. Also the slightly higher accuracy observed for BAYES_A compared to the other two methods on fat and protein percentage can support the above reported considerations.

Observed reliability accuracies of genomic predictions were similar or higher to those of traditional pedigree indexes in the case of Brown bulls but rather smaller, except for percentage of fat and protein milk yield, in the case of Simmental bulls. Even though genomic prediction have been reported to be more accurate than PI (De Los Campos et al., 2010; Olson et al., 2011; VanRaden et al., 2009), these are rather expected results, considering the limited size of the samples used in this study.

Obtained DGV accuracies GEBV reliablitiy are characterized by a relevant variation both within and between breeds. In particular, the Brown breed showed a higher variation in $R^2_{DGV,EBV}$ across traits compared to the Simmental. Differences in genomic accuracies between traits have

been reported in other papers (Hayes et al., 2009; Su et al., 2010; VanRaden et al., 2009) even though not of this magnitude. Moreover, it has to be remembered that most of literature deals with Holstein cattle. In any case, apart from the different genetic background of the considered traits, the sample size together with the wide range of birth year of bulls can be reasonably considered main causes of the present results. This consideration may explain the relevant reduction in accuracy for milk yield in the last scenario (REF:PRED 90:10) of the Brown bulls (Table 3). Actually this trait has been intensively selected across years and therefore the youngest 75 Brown bulls are very far from many REF animals both in terms of time and of genetic background(ora è il contrario). Therefore, PC or maker effects estimated in the REF population can be not adequate to predict their DGV GEBV. Actually, the random inclusion of some of the youngest animals in the REF data set results in an increase of accuracy in the yield traits (Table 4). Reasons for the different behavior of the Simmental breed (less variation between traits, higher values for milk yield) remain unclear less clear. A partial explanation could be found in the pattern of birth year of bulls, narrower compared with Brown. Moreover, the lower accuracy for fat percentage compared to Brown should be ascribed to the known fixation of the favorable mutation at the DGAT1 locus in the Italian Simmental.

CONCLUSIONS

Principal Component Analysis was effective in reducing the number of predictors needed for calculating direct genomic values genomic enhanced estimated breeding values for dairy traits in Brown and Simmental bulls. Such a reduction did not affect DGV and GEBV accuracy and allowed for a relevant decrease of calculation time. The obtained accuracies, although moderate to low mainly due to the size of the sample of animals considered, highlighted some differences between traits ad breeds. Results of the present work suggest the PC approach as a possible

alternative to the use of SNP genotypes for predicting DGV, especially for populations of limited size.

ACKNOWLEDGMENTS

Research funded by the Italian Ministry of Agriculture, grant SELMOL.

REFERENCES

Aulchenko, Y. S., D.-J. de Koning, and C. Haley. 2007. Genomewide Rapid Association Using Mixed Model and Regression: A Fast and Simple Method For Genomewide Pedigree-Based Quantitative Trait Loci Association Analysis. Genetics 177(1):577-585.

Boichard, D., V. Ducrocq, S. Fritz and J. J. Colleau, 2010 Where is dairy cattle breeding going? A vision of the future. Interbull Workshop on the Use of Genomic Information in Genetic evaluations. Paris, March 4-5, 2010

Bolormaa, S., J. E. Pryce, B. J. Hayes, and M. E. Goddard. 2010. Multivariate analysis of a genome-wide association study in dairy cattle. J. Dairy Sci. 93(8):3818-3833.

Bumb, B., 1982 Factor analysis and development. J. Develop. Econom. 11: 109-112.

Burden, R. L. and J.D. Faires. Numerical Analysis. Thomson Brooks/Cole. Belmont, CA, USA.

Cavalli-Sforza, L. L. and M. W. Feldman 2003. The application of molecular genetic approaches to the study of human evolution. Nat Genet.

Cohen-Zinder, M., E. Seroussi, D. M. Larkin, J. J. Loor, A. Everts-van der Wind, J. H. Lee, J. K. Drackley, M. R. Band, A. G. Hernandez, M. Shani, H. A. Lewin, J. I. Weller, and M. Ron. 2005. Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. Genome Res. 15(7):936-944.

Cole, J. B., P. M. VanRaden, J. R. O'Connell, C. P. Van Tassell, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and G. R. Wiggans. 2009. Distribution and Location of Genetic effects for Dairy traits (vol 92, pg 2931, 2009). J. Dairy Sci. 92(7):3542-3542.

de los Campos, G., D. Gianola, and D. B. Allison. 2010. Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat. Rev. Genet. 11, 880-886

de Roos, A. P. W., B. J. Hayes, and M. E. Goddard. 2009. Reliability of Genomic Predictions Across Multiple Populations. Genetics 183(4):1545-1553.

Ducrocq, V., and Liu, Z. 2009. Combining genomic and classical information in national BLUP evaluations. Interb. Bull. 40:172-177.

441 Ghiroldi, S., C. Nicoletti, E. Santus, A. Rossoni, and A. Bagnato. 2006. ITE: The new selection
442 index for the Italian Brown Swiss. 2005. Interb.l Bull.. 33:222-177.
443
444 Gianola, D., R. L. Fernando, and A. Stella. 2006. Genomic-Assisted Prediction of Genetic Value
445 With Semiparametric Procedures. Genetics 173(3):1761-1776.
446
447 Gianola, D. and J. B. C. H. M. van Kaam. 2008. Reproducing Kernel Hilbert Spaces Regression
448 Methods for Genomic Assisted Prediction of Quantitative Traits. Genetics 178(4):2289-2303.
449
450 Goddard, M. 2009. Genomic selection: prediction of accuracy and maximisation of long term
451 response. Genetica 136(2):245-257.
452
453 Gredler, B. ,H. Schwarzenbacher, C. Egger-Danner, C. Fuerst, R. Emmerling, and J. Sölkner. 2010.
454 Accuracy of genomic selection in dual purpose Fleckvieh cattle using three types of methods and
455 phenotypes. Proc. 9th World Congr. Genet. Appl. Livest. Prod. Article n. 0907
456
457 Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid,
458 P. Simon, R. Spelman, M. Georges, and R. Snell. 2002. Positional candidate cloning of a QTL in
459 dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on
460 milk yield and composition. Genome Res 12(2):222-231.
461
462 Guo, G., M. Lund, Y. Zhang, and G. Su. 2010. Comparison between genomic predictions using
463 daughter yield deviation and conventional estimated breeding value as response variables. J. Anim.
464 Breed. Genet. 127: 423–432.
465
466 Habier, D., J. Tetens, F. R. Seefried, P. Lichtner, and G. Thaller. 2010. The impact of genetic
467 relationship information on genomic breeding values in German Holstein cattle. Genet. Sel. Evol.
468 42.
469
470 Hayes, B. and M. E. Goddard. 2001. The distribution of the effects of genes affecting quantitative
471 traits in livestock. Genet. Sel. Evol. 33(3):209-229.
472
473 Hayes, B. J., P. J. Bowman, A. C. Chamberlain, K. Verbyla, and M. E. Goddard. 2009a. Accuracy
474 of genomic breeding values in multi-breed dairy cattle populations. Genet. Sel. Evol. 41.
475
476 Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009b. Invited review:
477 Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. 92(2):433-443.
478
479 Henderson, C. R. 1985. Best Linear Unbiased Prediction Using Relationship Matrices Derived from
480 Selected Base Populations. J. Dairy Sci. 68(2):443-448.
481
482 Jombart, T., D. Pontier, and A. B. Dufour. 2009. Genetic markers in the playground of multivariate
483 analysis. Heredity 102(4):330-341.
484
485 Konig, S., H. Simianer, and A. Willam. 2009. Economic evaluation of genomic breeding programs.
486 J. Dairy Sci. 92(1):382-391.
487
488 Lande, R. and R. Thompson. 1990. Efficiency of Marker-Assisted Selection in the Improvement of
489 Quantitative Traits. Genetics 124(3):743-756.
490

491 Legarra, A. and I. Misztal. 2008. Technical note: Computing strategies in genome-wide selection. J.
492 Dairy Sci. 91(1):360-366.
493
494 Long, N., D. Gianola, G. J. M. Rosa, K. A. Weigel, and S. Avendaño. 2007. Machine learning
495 classification procedure for selecting SNPs in genomic selection: application to early mortality in
496 broilers. J. Anim. Breed. Genet. 124(6):377-389.
497
498 Long, N., D. Gianola, G.J.M. Rosa and K.A. Weigel. 2011. Dimension reduction and variable
499 selection for genomic selection: application to predicting milk yield in Holsteins. J. Anim. Breed.
500 Gene.128.
501
502 Luan, T., J. A. Woolliams, S. Lien, M. Kent, M. Svendsen, and T. H. E. Meuwissen. 2009. The
503 Accuracy of Genomic Selection in Norwegian Red Cattle Assessed by Cross-Validation. Genetics
504 183(3):1119-1126.
505
506 Muir, W.M.. 2007. Comparison of genomic and traditional BLUP-estimated breeding value
507 accuracy and selection response under alternative trait and genomic parameters. J. Anim. Breed.
508 Genet. 124:342-355.
509
510
511 Macciotta, N. P. P., G. Gaspa, R. Steri, E. L. Nicolazzi, C. Dimauro, C. Pieramati, and A. Cappio-
512 Borlino. 2010. Using eigenvalues as variance priors in the prediction of genomic breeding values by
513 principal component analysis. J. Dairy Sci. 93(6):2765-2774.
514
515 Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using
516 genome-wide dense marker maps. Genetics 157(4):1819-1829.
517
518 Moser, G., M. Khatkar, B. Hayes, and H. Raadsma. 2010. Accuracy of direct genomic values in
519 Holstein bulls and cows using subsets of SNP markers. Genet. Sel. Evol. 42(1):37.
520
521 Moser, G., B. Tier, R. E. Crump, M. S. Khatkar, and H. W. Raadsma. 2009. A comparison of five
522 methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genet.
523 Sel. Evol. 41.
524
525 Olson, K.M., P.M. VanRaden, M.E. Tooker, and T.A. Cooper. 2011. Differences among methods to
526 validate genomic evaluations for dairy cattle. J.Dairy. Sci. 94:2613-2620.
527
528 Paschou, P., E. Ziv, E. G. Burchard, S. Choudhry, W. Rodriguez-Cintron, M. W. Mahoney, and P.
529 Drineas. 2007. PCA-Correlated SNPs for Structure Identification in Worldwide Human
530 Populations. Plos Genet 3(9):e160.
531
532 Pausch, H., K. Flisikowski, S. Jung,R.Emmerling, C. Edel, K.U. Gotz, and R. Fries. 2011. Genome-
533 Wide Association Study Identifies Two Major Loci Affecting Calving Ease and Growth-Related
534 Traits in Cattle. Genetics 187: 289–297
535
536 Patterson, N., A. L. Price, and D. Reich. 2006. Population Structure and Eigenanalysis. Plos Genet
537 2(12):e190.
538

539 Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. 2006.
540 Principal components analysis corrects for stratification in genome-wide association studies. Nat
541 Genet 38(8):904-909.
542
543
544 Schaeffer, L. R. 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed.
545 Genet. 123(4):218-223.
546
547 Solberg, T. R., A. K. Sonesson, J. A. Woolliams, and T. H. E. Meuwissen. 2009. Reducing
548 dimensionality for prediction of genome-wide breeding values. Genet. Sel. Evol. 41:-.
549
550 Su, G., B. Guldbrandtsen, V. R. Gregersen, and M. S. Lund. 2010. Preliminary investigation on
551 reliability of genomic estimated breeding values in the Danish Holstein population. J. Dairy Sci.
552 93(3):1175-1183.
553
554
555 Van Tassell, C. P., T. P. L. Smith, L. K. Matukumalli, J. F. Taylor, R. D. Schnabel, C. T. Lawley,
556 C. D. Haudenschild, S. S. Moore, W. C. Warren, and T. S. Sonstegard. 2008. SNP discovery and
557 allele frequency estimation by deep sequencing of reduced representation libraries. Nat Meth
558 5(3):247-252.
559
560 VanRaden, P. M. 2008. Efficient Methods to Compute Genomic Predictions. J. Dairy Sci.
561 91(11):4414-4423.
562
563 VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor,
564 and F. S. Schenkel. 2009. Invited review: Reliability of genomic predictions for North American
565 Holstein bulls. J. Dairy Sci. 92(1):16-24.
566
567 VanRaden, P. M. and P. G. Sullivan. 2010. International genomic evaluation methods for dairy
568 cattle. Genet. Sel. Evol. 42:-.
569
570 VanRaden, P. M., J.R. O'Connell, G. R. Wiggans, K. Weigel. 2011. Genomic evaluations with
571 many more genotypes. Genet. Sel. Evol. 43:10.
572
573 Vazquez, A. I., G. J. M. Rosa, K. A. Weigel, G. de los Campos, D. Gianola, and D. B. Allison.
574 2010. Predictive ability of subsets of single nucleotide polymorphisms with and without parent
575 average in US Holsteins. J. Dairy Sci. 93(12):5942-5949.
576
577 Visscher, P. M., S. Macgregor, B. Benyamin, G. Zhu, S. Gordon, S. Medland, W. G. Hill, J. J.
578 Hottenga, G. Willemsen, D. I. Boomsma, Y. Z. Liu, H. W. Deng, G. W. Montgomery, and N. G.
579 Martin. 2007. Genome partitioning of genetic variation for height from 11,214 sibling pairs. Am J
580 Hum Genet 81(5):1104-1110.
581
582 Wiggans, G. R., T. S. Sonstegard, P. M. Vanraden, L. K. Matukumalli, R. D. Schnabel, J. F. Taylor,
583 F. S. Schenkel, and C. P. Van Tassell. 2009. Selection of single-nucleotide polymorphisms and
584 quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. J.
585 Dairy Sci. 92(7):3431-3436.
586

587 Yang, J., B. Benyamin, B.P McEvoy, S.Gordon, A.K. Henders, D.R. Nyholt, P.A. Madden, A C.
588 Heath, N.G. Martin, G.W. Montgomery, et al. 2010. Common SNPs explain a large proportion of
589 the heritability for human height. Nat. Genet. 42: 565-569.
590

591

592 **Table 1.** Number of animals and markers discarded in the different edit steps.

Breed	Repeated[1]	Mendelian Inheritance[2]	Missing[3]	MAF[4]	HW[5]	Final dataset DRG
		Animals				
Brown	17	3	6			634
Simmental	6	2	6			469
		SNP markers				
Brown		23	1,118	15,046	560	37,254
Simmental		21	999	12.215	587	40,179

593

594 [1]Number of animals genotyped twice to check genotyping quality

595 [2]SNP that showed Mendelian conflicts in more than 2.5% father-sons pairs; animals that showed

596 more than 2,000 Mendelian conflicts .

597 [3]Animals with more than 1,000 missing genotypes; SNP with more than 2.5% missing genotypes

598 [4] SNP with a minor allele frequency lower than 0.05.

599 [5] SNP that deviate significantly ($P<0.01$) from Hardy Weinberg equilibrium.

600

Table 2. Pearson correlations (X100) between direct genomic values and polygenic estimated breeding values , for different estimation methods for both Simmental and Brown datasets

	PC-BLUP		SNP-BLUP		BAYES A	
	BROWN	SIMENT	BROWN	SIMENT	BROWN	SIMENT
Trait			*Ref:Pred 70:30*			
Milk yield	17.4	31.0	4.9	36.5	14.6	37.7
Fat yield	26.1	24.5	16.5	27.6	29.9	28.6
Protein yield	15.7	22.8	6.8	28.4	16.9	29.9
SCC	23.9	1.4	13.5	15.2	25.4	16.4
Fat percentage	40.5	13.8	18.3	18.4	45.1	18.2
Protein percentage	47.4	33.2	24.2	33.5	46.5	35.1
			Ref:Pred 80:20			
Milk yield	18.1	44.6	6.6	41.2	17.5	42.3
Fat yield	26.4	34.8	21.9	27.4	31.1	28.8
Protein yield	18.2	40.9	11.5	30.9	22.1	33.6
SCC	31.7	9.3	25.9	15.1	32.5	17.2
Fat percentage	40.7	3.3	28.2	9.1	42.5	6.2
Protein percentage	42.4	30.8	25.9	31.0	38.8	31.9
			Ref:Pred 90:10			
Milk yield	28.9	51.4	14.8	42.7	16.8	45.5
Fat yield	43.5	40.3	35.9	34.9	41.8	36.0
Protein yield	40.7	48.7	26.3	35.5	32.7	38.1
SCC	4.2	7.3	11.6	17.3	38.7	11.8
Fat percentage	35.4	12.4	19.8	8.5	34.9	11.9
Protein percentage	53.2	21.3	28.4	25.9	40.8	21.5

623

624

625 **Table 3.** Average (standard deviations in brackets) Pearson correlations between predicted direct
626 genomic breeding values and polygenic breeding values in the two breeds using Principal
627 component scores as predictors when prediction population (30% of the whole data set) is created
628 by randomly picking up animals from the 50% of youngest bulls

Trait	Brown	Simmental
Milk yield	27.3 (2.5)	46.3 (2.3)
Fat yield	32.7 (1.7)	39.1 (2.8)
Protein yield	33.2 (3.0)	43.6 (4.0)
Fat percentage	43.7 (3.2)	18.0 (4.7)
Protein percentage	49.2 (3.8)	30.8 (4.3)
SCC	34.0 (6.9)	25.4 (7.0)

629

630

631

632 **Table 4.** Squared correlations between genomic enhanced breeding values obtained using principal
633 component scores (PC_BLUP) as predictors, or SNP genotypes with a BLUP (SNP_BLUP) or
634 Bayesd A (BAYES_A) methods, or pedigree indexes (PI) and polygenic estimated breeding values,
635 for different scenarios in the Brown breed.

	Estimation method			
Trait	PC_BLUP	SNP_BLUP	BAYES_A	PI
	Ref:Pred 70:30			
Milk yield	4.5	1.6	3.6	4.6
Fat yield	9.3	6.0	9.9	5.7
Protein yield	2.7	1.1	2.5	3.5
SCC	13.9	13.2	13.4	12.5
Fat percentage	35.1	30.4	35.2	25.6
Protein percentage	38.4	30.5	34.9	29.8
	Ref:Pred 80:20			
Milk yield	9.0	4.6	7.8	8.6
Fat yield	9.7	8.1	10.4	6.3
Protein yield	2.4	1.0	2.3	2.2
SCC	11.7	11.2	10.9	9.7
Fat percentage	38.5	34.4	36.7	26.7
Protein percentage	34.2	28.8	30.6	24.5
	Ref:Pred 90:10			
Milk yield	12.3	7.1	6.6	6.6
Fat yield	22.9	19.2	18.4	8.3
Protein yield	12.6	3.5	2.9	0.4
SCC	21.0	22.0	19.8	20.9
Fat percentage	36.7	34.1	34.5	28.4
Protein percentage	37.6	26.3	27.1	20.4

636
637

638 **Table 4.** Squared correlations between genomic enhanced breeding values obtained using principal
639 component scores (PC_BLUP) as predictors, or SNP genotypes with a BLUP (SNP_BLUP) or
640 Bayesd A (BAYES_A) methods, or pedigree indexes (PI) and polygenic estimated breeding values,
641 for different scenarios in the Simmental breed.

Trait	Estimation method			
	PC_BLUP	SNP_BLUP	BAYES_A	PI
	Ref:Pred 70:30			
Milk yield	36.6	35.4	35.8	34.5
Fat yield	34.3	33.8	33.9	33.3
Protein yield	35.3	34.1	34.4	34.1
SCC	20.1	20.4	20.3	20.5
Fat percentage	14.8	15.0	14.7	15.4
Protein percentage	20.2	19.0	19.4	21.0
	Ref:Pred 70:30			
Milk yield	36.7	35.3	35.7	33.1
Fat yield	31.2	30.0	30.3	28.8
Protein yield	33.0	30.6	31.0	30.5
SCC	20.3	20.5	20.6	20.6
Fat percentage	12.7	14.9	14.1	15.9
Protein percentage	17.9	16.5	17.4	16.9
	Ref:Pred 70:30			
Milk yield	36.6	30.4	31.8	24.8
Fat yield	29.4	27.3	27.8	23.4
Protein yield	32.7	24.0	25.1	20.5
SCC	18.2	18.3	17.8	18.2
Fat percentage	5.2	6.0	5.5	7.0
Protein percentage	11.9	15.2	13.3	15.0

642

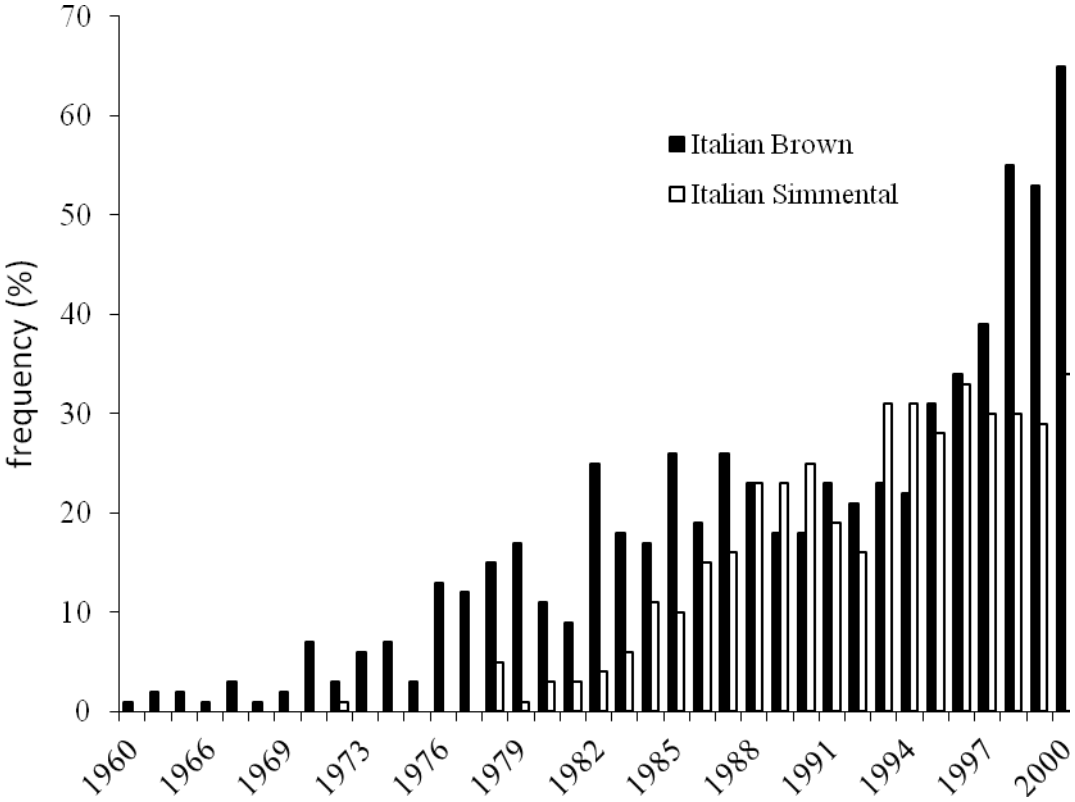
643

644 **Table 6.** Regression coefficients of polygenic breeding values on genomic enhanced breeding
645 values ($b_{EBV,GEBV}$) or PI ($b_{EBV,PI}$) for some dairy traits in Brown and Simmental prediction animals
646 using principal components scores (PC_BLUP), SNP genotypes (SNP_BLUP) or Bayes
647 (BAYES_A) estimation method.

		BROWN			SIMMENTAL		
Trait	Method	70:30	80:20	90:10	70:30	80:20	90:10
Milk yield	PC_BLUP	0.49	0.66	0.86	1.09	1.00	0.96
	SNP_BLUP	0.26	0.45	0.59	1.12	1.10	1.01
	BAYES_ A	0.47	0.71	0.70	1.12	1.06	1.04
	PA	0.31	0.44	0.41	0.91	0.88	0.73
Fat yield	PC_BLUP	0.80	0.83	1.26	1.05	1.06	1.20
	SNP_BLUP	0.56	0.66	1.00	1.09	1.11	1.38
	BAYES_ A	0.93	0.99	1.34	1.09	1.11	1.38
	PA	0.39	0.43	0.48	0.93	0.94	1.05
Protein yield	PC_BLUP	0.42	0.41	1.01	1.00	0.99	1.10
	SNP_BLUP	0.22	0.23	0.47	1.02	0.99	1.04
	BAYES_ A	0.43	0.44	0.62	1.04	0.99	1.07
	PA	0.29	0.25	0.13	0.87	0.85	0.79
SCS	PC_BLUP	2.27	2.17	2.53	0.73	0.73	0.83
	SNP_BLUP	1.95	1.86	2.28	0.78	0.77	0.88
	BAYES_ A	2.28	2.15	2.57	0.78	0.77	0.87
	PA	0.80	0.73	0.94	0.73	0.72	0.81
Fat percentage	PC_BLUP	1.33	1.35	1.48	0.59	0.64	0.47
	SNP_BLUP	1.20	1.31	1.29	0.65	0.65	0.59
	BAYES_ A	1.46	1.54	1.46	0.64	0.64	0.56
	PA	0.78	0.80	0.80	0.53	0.54	0.46
Protein percentage	PC_BLUP	1.29	1.18	1.45	0.88	0.93	0.72
	SNP_BLUP	1.13	1.18	1.21	0.96	0.88	0.89
	BAYES_A	1.33	1.32	1.32	0.96	0.91	0.85
	PA	0.81	0.76	0.77	0.83	0.73	0.68

648

649



650

651

652

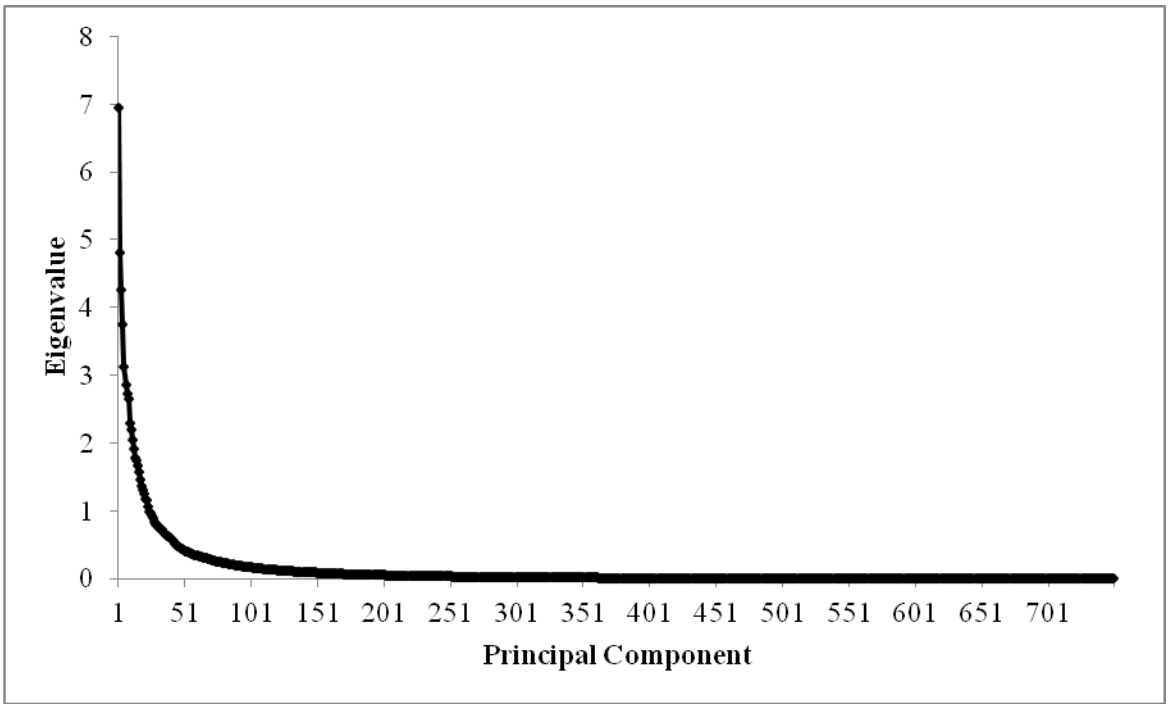
653

654

655 **FIGURE 1**. Distribution of number of bulls across year of birth.

656

657



658

659 **FIGURE 2.** Pattern of the proportion of variance (%) accounted for by each successive principal component

660 extracted from the correlation matrix of SNP markers for the chromosome six in the Brown breed.

661

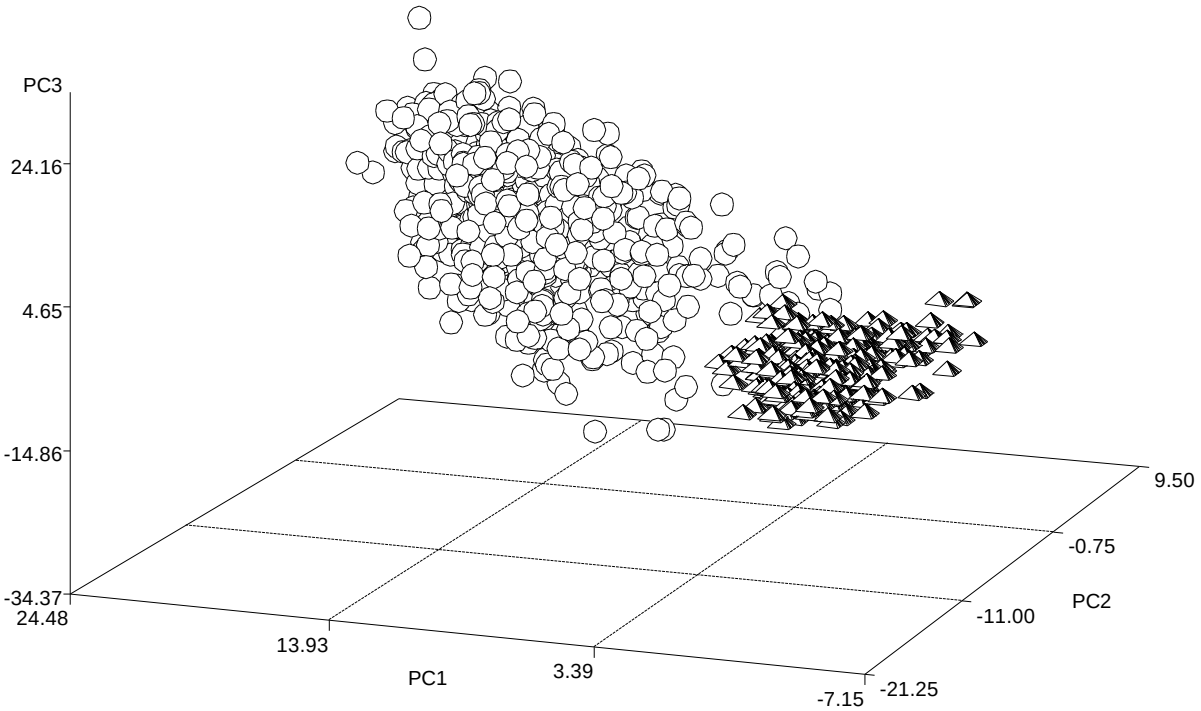


FIGURE 3. Plot of the individual scores of the first three principal components (PC1, PC2 and PC3) extracted from chromosome six in the two breeds (Circles=Brown; Pyramids=Simmental)..